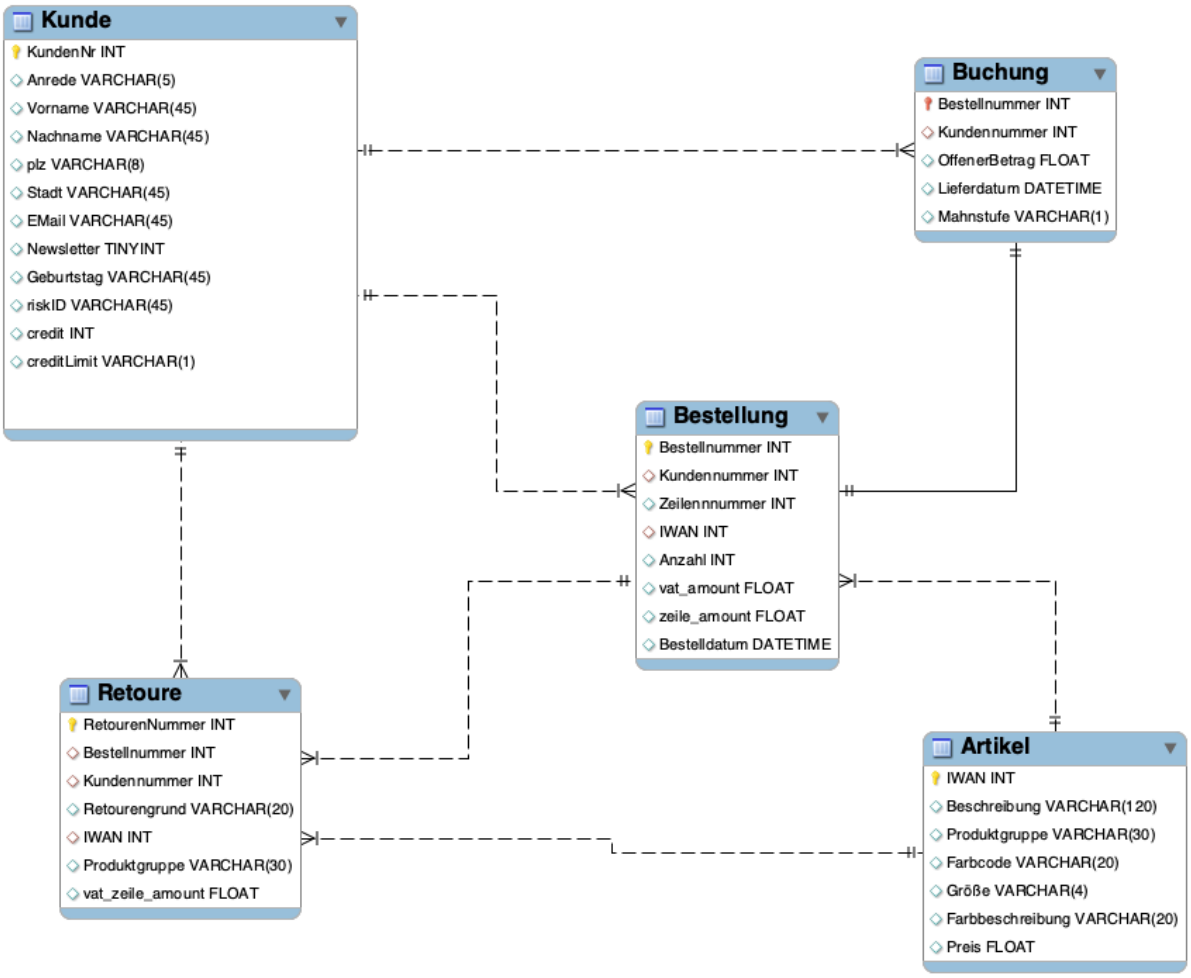
Übung A1: ETL mit Polars

- Dokumentation zu der Bibliothek Polars

Aufgabenstellung

Beantworten Sie folgende Fragestellungen bzw. führen Sie folgende Aufgaben bezüglich des Webshops mittels der Bibliothek Polars Version 1.x (stable) aus. Das ER-Diagramm des Webshops, Information zu den Daten finden Sie weiter unten in der Aufgabenstellung.

a) Lesen Sie die Daten für die Entitäten Kunde, Buchung und Artikel, Bestellung, Retoure gemäß des Datenmodells ein, d.h. passen Sie die Namen der Attribute an, reduzieren Sie die Daten auf die Attribute des Datenmodells, beachten Sie die Datentypen.
b) Geben Sie die 5 jüngsten Kunden aus.
c) Entfernen Sie alle Kunden aus "Kunde", die vor 1930 geboren sind.
d) Ergänzen Sie das Dataframe durch ein Attribut 'Bundesland', in dem das Bundesland zur PLZ gespeichert sein soll. Die Bundesländer entnehmen Sie bitte aus der Datei `plz_mapping.txt`.
e) Leider wird in dem Webshop i.d.R. als Gast ohne Kundenkonto bestellt. Daher existieren viele Kunden mit nahezu gleichen Stammdaten (Duplikate), aber mit unterschiedlicher KundenNr. Alle Datensätze mit gleichem Wert im Attribut 'riskId' sollen als ein eindeutiger Kunde in einem neuen DataFrame 'uniqueCustomers' zusammengeführt werden. Als KundenNr. und alle anderen Stammdaten sollen die Einträge des ersten Vorkommens übernommen werden. Das DataFrame 'uniqueCustomers' soll persistiert und für die Beantwortung der noch folgenden Punkte verwendet werden.
f) Ergänzen Sie das Dataframe Artikel um ein weiteres Attribut "Länge", dass die Anzahl Buchstaben des Attributs "Beschreibung" beziffert.
g) Erstellen Sie je ein Balkendiagramm, das die Verteilung des Attributs "Länge" und "Preis" je Produktgruppe des Dataframes Artikel zeigt. Nutzen Sie bitte die Methoden im Namespace df.plot.
h) Erstellen Sie ein Balkendiagramm, das die Verteilung der Kunden je Bundesland und je Geschlecht visualisiert. Die Verteilung soll absteigend sortiert nach Häufigkeit der Kunden je Bundesland sein. Nutzen Sie bitte die Methoden im Namespace df.plot.
i) Kategorisieren Sie das Attribut Preis in drei "gleichgroße" Kategorien, jede Kategorie soll "gleich" viele Artikel abdecken. Gleiche Preise sollen der gleichen Kategorie zugeordnet sein. Fügen Sie das kategorisierte Attribut als "Preiskategorie" dem Dataframe "Artikel" hinzu.
j) Gruppieren Sie die Umsätze des Jahres 2010 je Monat und stellen Sie diese in einem Balkendiagramm dar. Die Spalte "*Bestelldatum"* ist für die Feststellung des Bestelldatums zu verwenden. Berücksichtigen Sie auch etwaige Retouren von Bestellungen aus dem genannten Zeitraum. Jede Bestellposition (Zeilennummer) wird bei den Retouren einzeln unter der gleichen RetourenNummer gelistet. Es ist daher durchaus möglich, dass es mehrere Retouren zu einer Bestellung gibt.
k) Gibt es Produktgruppen, die bei den Kunden zu erhöhten Mahnstufen führen? Zeigen Sie die durchschnittliche Mahnstufe je Produktgruppe über die gekauften Artikel, Retouren sollen nicht berücksichtigt werden. Geben Sie die 10 Produktgruppen mit den höchsten durchschnittlichen Mahnstufen aus.
l) Was ist die durchschnittliche Anzahl an Bestellpositionen (Zeilennummern) und der durchschnittliche Gesamtbetrag einer Bestellung über alle Kunden? *(Hinweis: Es kann durchaus mehrere Bestellpositionen zu einer Bestellung geben, jede Bestellung ist eindeutig durch die Spalte Bestellnummer).*
m) Ermitteln Sie die Anzahl der Bestellpositionen und die Anzahl der retournierten Positionen je Bundesland. Die Kundenadresse ist hier als Ort der Bestellung anzunehmen. Bestimmen Sie anschließend die Retourenquote (= Anzahl an Retourenpositionen / Anzahl an Bestellpositionen)



Die Input-Dateien finden Sie hier:

https://cloud.th-luebeck.de/index.php/s/DD2joB2ZWNiNi39.

Folgend finden Sie eine Beschreibung der Input-Dateien:

iw_customer		Kunden
Spaltenname	**Datentyp**	**Inhalt**
owner	varchar	Shop-/Mandatenkennung
customerNo	varchar	Kundennummer
salutation	varchar	Anrede
firstname	varchar	Vorname (anonymisiert)
surname	varchar	Nachname (anonymisiert)
postcode	varchar	Postleitzahl
city	varchar	Wohnort
street	varchar	Straße (anonymisiert)
eMail	varchar	E-Mail (anonymisiert)
newsletter	varchar	Newsletter (1 = Ja, 0 = Nein)
birthdate	datetime	Geburtsdatum
riskID	varchar	ID der Bonitätsprüfung (dient der Identifikation eines Kunden)
credit	numeric	Höhe des zulässigen Kredits
creditLimit	varchar	1 = hat Kredit, 2 = kein Kredit

iw_sales		Bestellungen
Spaltenname	**Datentyp**	**Inhalt**
owner	varchar	Shop-/Mandatenkennung
line_No	numeric	Zeilennummer der Rechnung
orderNo	varchar	Bestellnummer
customerNo	varchar	Kundennummer
type	numeric	2 = Artikel, 1 = Versand
IWAN	numeric	Eindeutige Artikelnummer wie EAN (5000 = Versand)
quantity	numeric	Anzahl der Artikel
amount	money	Nettopreis des Artikels
vat_amount	money	Preis inkl. MwSt.
line_amount	money	Summe der Zeile inkl. MwSt.
VATpercent	varchar	Mehrwertsteuersatz
bill_customerNo	varchar	Kundennummer des Rechnungsempfängers
orderDate	datetime	Bestelldatum
postingDate	datetime	Verarbeitungsdatum

iw_article		Artikel
Spaltenname	**Datentyp**	**Inhalt**
owner	varchar	Shop-/Mandatenkennung
IWAN	numeric	Eindeutige Artikelnummer wie EAN
article_No	varchar	Interne Artikelnummer
description	varchar	Artikelbeschreibung
unitPrice	money	Stückpreis
deftime	datetime	Zeitstempel Artikel gelistet
modtime	datetime	Zeitstempel Artikel zuletzt bearbeitet
seasonCode	varchar	Saison-Code
productGroup	varchar	Produktgruppen-Code
colorCode	varchar	Farb-Code
colorDescription	varchar	Farbbeschreibung
size	varchar	Größe
articleOnline	varchar	1 = online, 0 = offline

iw_payment		Buchungen
Spaltenname	**Datentyp**	**Inhalt**
owner	varchar	Shop-/Mandatenkennung
orderNo	varchar	Bestellnummer
customerNo	varchar	Kundennummer
outstandingAmount	money	Offener Betrag (Rechnungsbeträge > 0, Zahlungen und Retouren < 0)
postingDate	datetime	Bearbeitungs-, Lieferdatum
dueDate	datetime	Zahlungsziel
closedAccountDate	datetime	Zeitstempel Konto geschlossen
openAccount	varchar	0 = geschlossen, 1 = offen
dunningLevel	varchar	Mahnstufen 1 bis 4 (0 = keine Mahnstufe)

iw_return_header		Retouren
Spaltenname	**Datentyp**	**Inhalt**
owner	varchar	Shop-/Mandatenkennung
returnNo	varchar	Retourennummer
orderNo	varchar	Bestellnummer
paymentCode	varchar	Zahlungsart
returnType	varchar	Retourentyp
shippingAgent	varchar	Versender
customerNo	varchar	Kundennummer
bill_customerNo	varchar	Kundennummer des Rechnungsempfängers
shipmentDate	datetime	Versanddatum
postingDate	datetime	Bearbeitungsdatum

iw_return_line		Retouren-Positionen
Spaltenname	**Datentyp**	**Inhalt**
owner	varchar	Shop-/Mandatenkennung
returnNo	varchar	Retourennummer
customerNo	varchar	Kundennummer
bill_customerNo	varchar	Kundennummer des Rechnungsempfängers
quantity	numeric	Artikelanzahl
unitPrice	money	Stückpreis
IWAN	varchar	Eindeutige Artikelnummer wie EAN
type	numeric	2 = Artikel, 1 = Versand
returnReason	varchar	Retourengrund
productGroup	varchar	Produktgruppe
vat_line_amount	money	Summe der Zeile inkl. MwSt.
line_amount	money	Summe der Zeile ohne MwSt.
shipmentDate	datetime	Versanddatum
postingDate	datetime	Bearbeitungsdatum

iw_code_reason		Codes
Spaltenname	**Datentyp**	**Inhalt**
owner	varchar	Shop-/Mandatenkennung
Type	varchar	returnReason/returnType/payment
Code	varchar	Code
Reason	varchar	Klartext

Laden Sie den erstellten Quelltext, e.g. Jupyter Notebooks, Python Files etc., als Zip-Archiv in den Lernraum hoch.